A Strategy for Characterization of Single Nucleotide Polymorphisms in a Reference Material

NIST
National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

Email:
Kevin.Kiesler@nist.gov

Kevin M. Kiesler, Katherine B. Gettings, and Peter M. Vallone
U.S. National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA

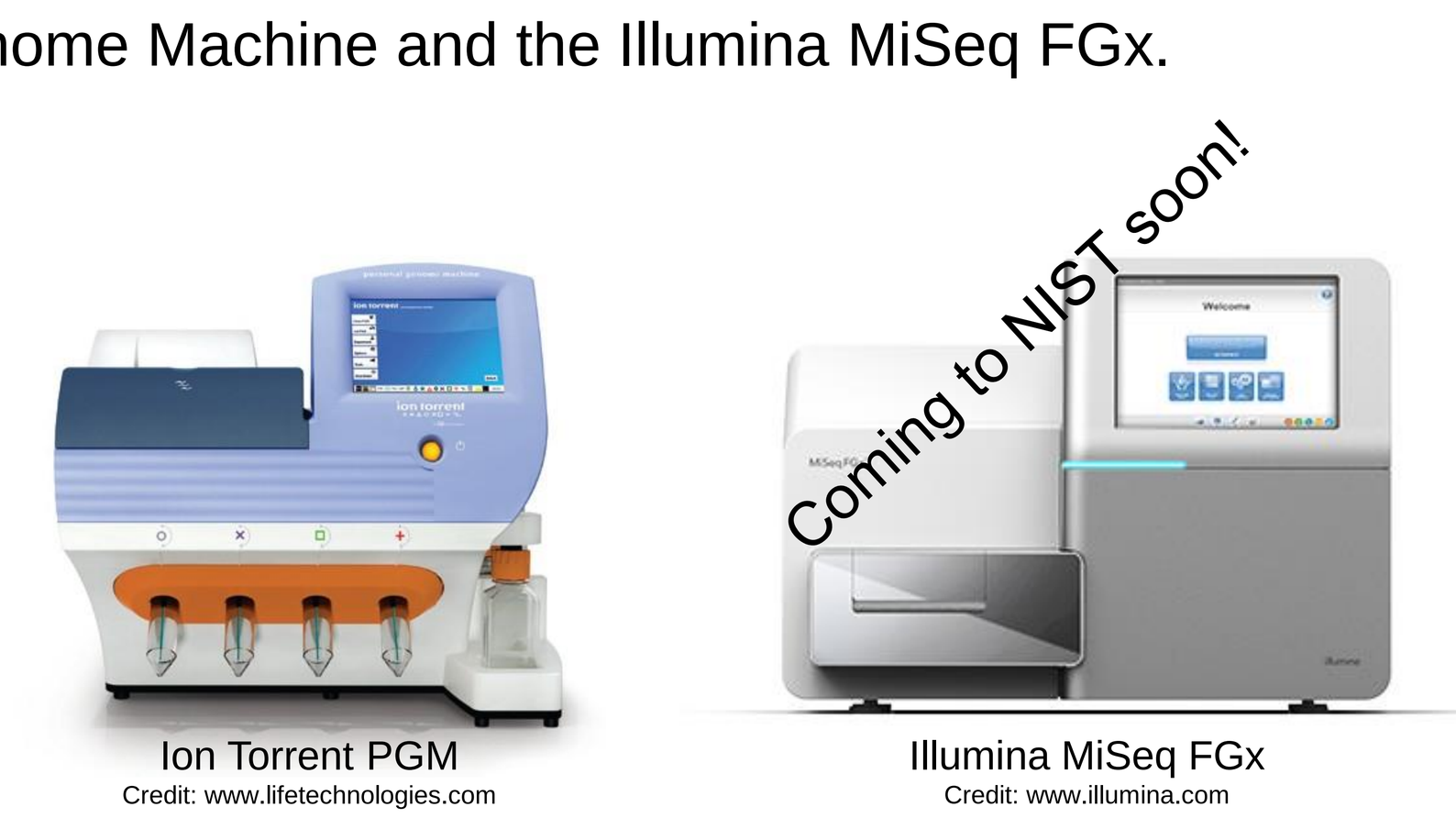
iSFG

P-192

The advent and adoption of next generation sequencing (NGS) is enabling analysis of single nucleotide polymorphisms (SNPs) at an unprecedented scale, limited primarily by multiplexing during the PCR amplification based enrichment step used for forensic applications. Since only a single nucleotide is assayed, PCR primers may be designed to generate small amplicons, making SNP markers well-suited to forensic DNA typing. Carefully selected panels of SNP markers have been previously established for forensic applications such as one-to-one matching, estimating biogeographical ancestry, and predicting externally observable phenotype [1,2,3,4,5,6]. To support the implementation of SNPs in forensic DNA analysis, NIST will examine the HID-Ion AmpliSeq Identity Panel and the HID-Ion AmpliSeq Ancestry Panel for the Ion Torrent Personal Genome Machine (PGM) and the Illumina ForenSeq DNA Signature kit for the Illumina MiSeq FGx. In total, over 300 SNP markers will be typed with approximately 40 % of loci being represented in more than one multiplex. A strategy combining NGS on orthogonal platforms and Sanger sequencing for characterizing the SNP markers for varying levels of confidence is presented herein. The outcome will be to report only the SNP allele calls, analogous to the mitochondrial sequence variants in NIST Standard Reference Material (SRM) 2392, and *not* the subsequent application of the information (e.g. prediction of ancestry or phenotype). Additionally, we demonstrate the optimization of automated sequencing template preparation using the Ion Chef from Ion Torrent to create an efficient workflow suitable to the forensic DNA laboratory.

Specific Aim: To devise an efficient methodology for characterization of SNP genotype values in a reference material. Using NGS as a primary measurement method, concordance between NGS platforms may be used to obtain high confidence SNP genotype values without exhaustive Sanger sequencing. SNP genotypes discordant across platforms may be disambiguated with Sanger sequencing. Evaluation of Mendelian inheritance patterns in family trios may be used as an additional confirmatory strategy.

Next Generation Sequencing Platforms:
Two platforms with commercial SNP genotyping panels intended for forensic DNA analysis are to be used in this characterization schema: the Ion Torrent Personal Genome Machine and the Illumina MiSeq FGx.



Ion Torrent PGM

HID-Ion AmpliSeq Identity Panel
90 Identity Informative SNPs (IISNPs) [1,2]
34 Lineage Informative SNPs (LISNPs) [3]

HID-Ion AmpliSeq Ancestry Panel
165 Ancestry Informative SNPs (AISNPs) [4,5]

Illumina MiSeq FGx

ForenSeq Signature Assay Kit - 172 SNP markers:
94 IISNPs [1,2]
56 AISNPs [5]
22 Phenotype Informative SNPs (PISNPs) [6]

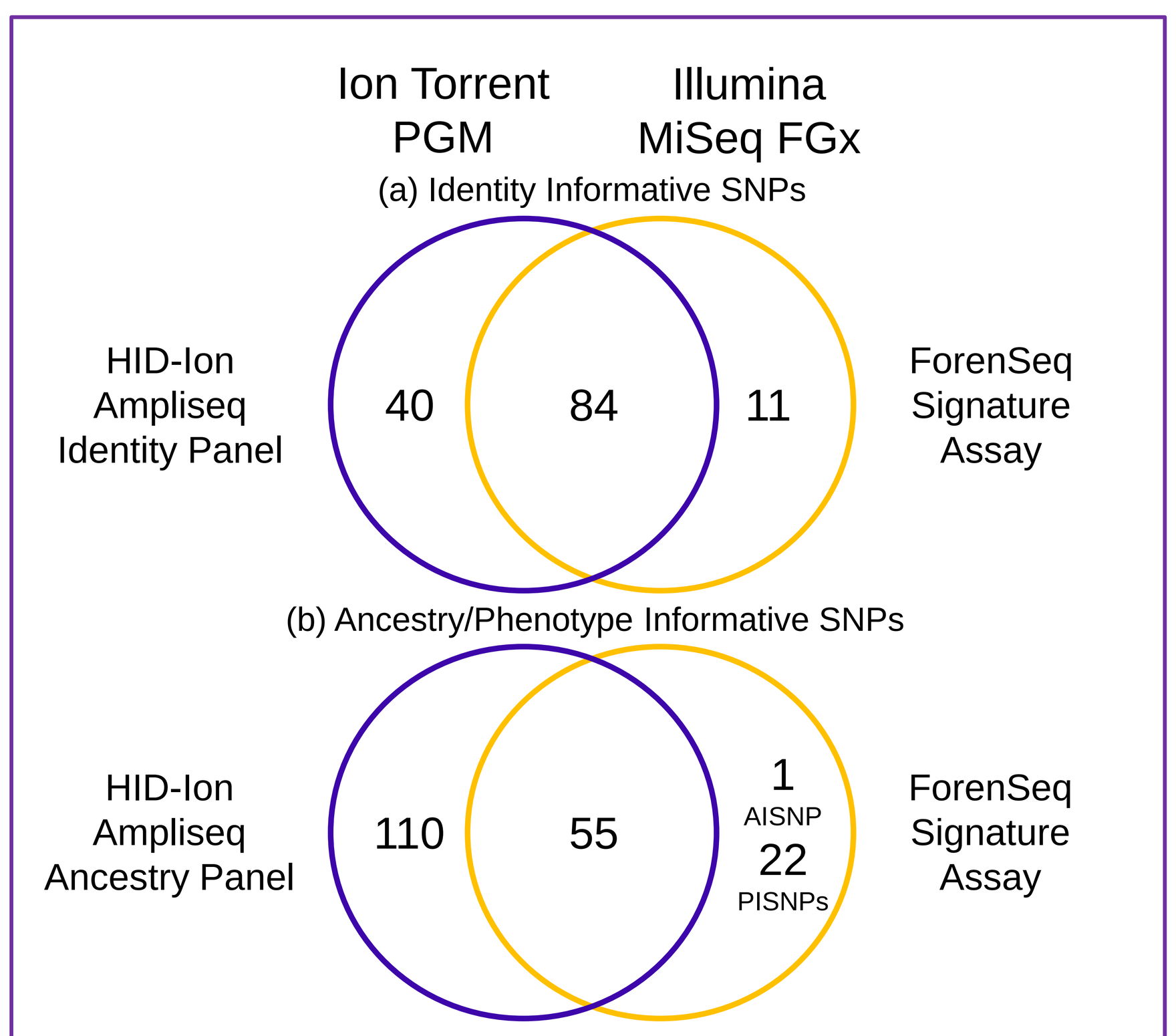


Figure 1: Venn diagrams of the SNP loci in Ion Torrent and Illumina NGS multiplexes with (a) overlapping identity informative markers (IISNPs) and (b) ancestry/phenotype informative markers (AISNPs). The Ion Torrent panels do not include any Phenotype Informative SNPs. There are 139 loci in common (out of a total 323) between the two platforms.

Methods:
Nine single-source DNA samples have been selected as candidate components for a reference material.

Name	Description
RM 1	Female
RM 2	Male
RM 3	Male
RM 4	Female
RM 5	Male
RM 6	Female
RM 7	Family Trio, Son
RM 8	Family Trio, Father
RM 9	Family Trio, Mother

- Typed on PGM Identity Panel & Ancestry Panel
- DNA input ≈ 1 ng
- Three replicate amplifications per sample
- Used library protocol for PGM HID SNPs
- ISP templating & chip loading performed on Ion Chef
- 31 samples in barcoded library pool
  - 9 candidate RMs + 1 control in triplicate
  - One negative control
- Sequenced each panel on one Ion 318 Chip

References:
1. Musgrave-Brown E., Ballard D., Balogh, K. *et al.*: Forensic validation of the SNP*for*ID 52-plex assay. *Forensic Science International: Genetics*. 2007; 1, 186–190.
2. Pakstis A, Speed W, Fang R, Hyland F, Furtado M, Kidd J, and Kidd K: SNPs for a universal individual identification panel. *Human Genetics*. 2010; 127(3), 315–324.
3. Karafet TM, Mendez FL, Meilerman M, Underhill P, Zegura S, and Hammer M.: New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Research*. 2008; 18(5), 830-838.
4. Nassir R, Kosoy R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM and Seldin MF: An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. *BMC Genetics*. 2009; 10(1), 39.
5. Kidd K, Speed W, Pakstis A, Furtado M, Fang R, Madbouly A, Maiers M, Middha M, Friedlander F, and Kidd J: Progress toward and efficient panel of SNPs for ancestry inference. *Forensic Science International: Genetics*. 2014; 10, 23-32.
6. Walsh S, Liu F, Wollstein A, Kovatsi L, Ralf A, Kosiniak-Kamysz A, Branicki W and Kayser M: The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Science International: Genetics*. 2013; 7, 98–115.
7. Seo S, King J, Warhsauer D, Davis C, Ge J, and Budowle B: Single nucleotide polymorphism typing with massively parallel sequencing for human identification. *Int J Legal Med*. 2013; 127, 1079-1086.

Financial disclosure: This work was supported by funding from the FBI Laboratory and Biometrics Center of Excellence: Forensic DNA Typing as a Biometric tool.

NIST Levels of Confidence for Certification

	Description	Proposed Method
Certified	NIST has the highest confidence in its accuracy. All known or suspected sources of bias have been investigated or taken into account.	Sanger Sequencing + NGS
Reference	A high-confidence estimate of the true value but where all possible sources of bias have not been fully investigated by NIST.	NGS + NGS Two Platform Concordance
Informational	Data that may be of interest and use to the SRM user, but insufficient information is available to access the confidence of the assignment.	NGS on one platform

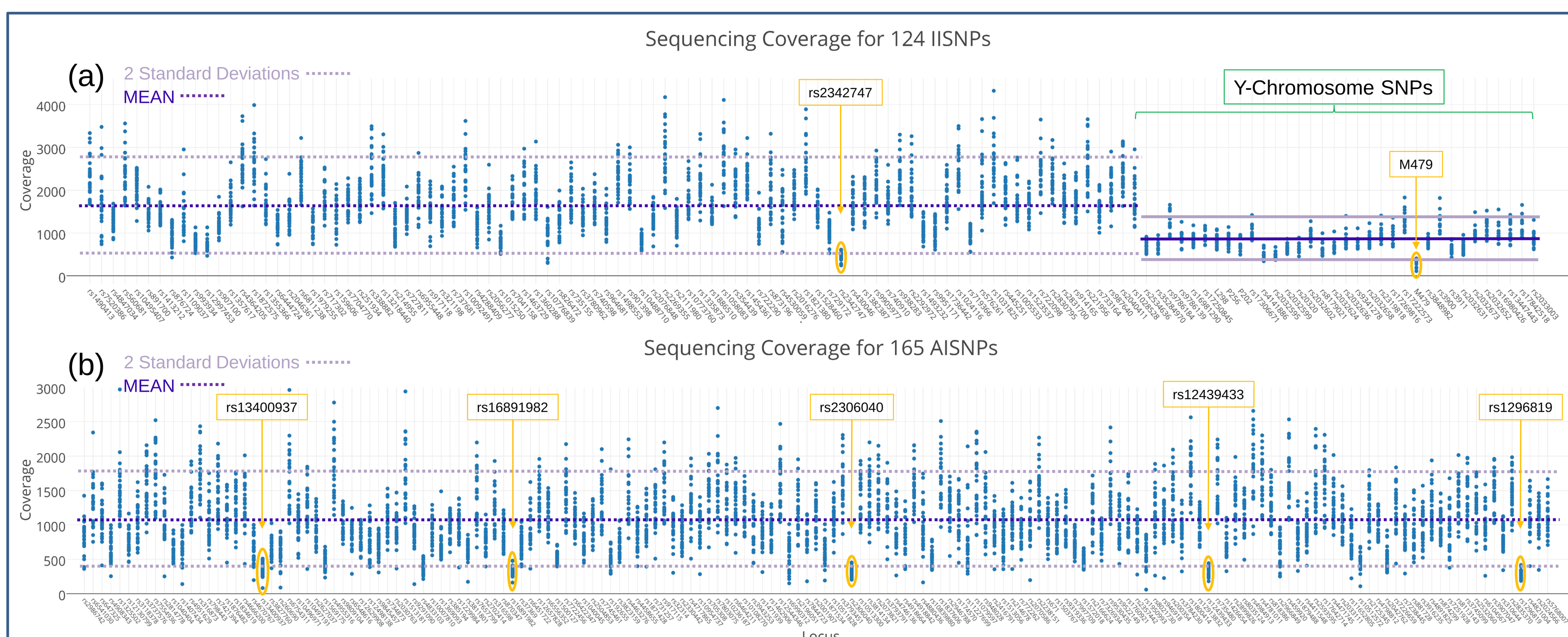


Figure 2: Sequencing coverage of nine candidate reference materials run in triplicate plotted for each SNP locus. (a) Identity Panel SNPs are shown with the 34 Y-chromosome LISNPs at the far right representing five male samples. (b) Ancestry Panel SNPs are shown. Loci with mean coverage below two standard deviations of the mean coverage for all loci are circled.

Indicators of Low Confidence Genotypes

Sequencing coverage: The Ion AmpliSeq Library Preparation for Human Identification Applications (Pub. No. MAN0010640 Rev. A.0) protocol suggests a minimum sequencing coverage of 300 X for autosomal SNPs (150 X for Y-SNPs). To achieve this minimum coverage level for all loci in a panel, an average coverage depth of 738 X is recommended for the Identity Panel and 594 X for the Ancestry Panel. These recommended values allow for locus-to-locus variations in sequencing coverage balance. This translates to a maximum of 77 samples in a library pool when using an Ion 318 Chip for the Identity Panel or 59 samples when running the Ancestry Panel, assuming 80 % chip loading and 60 % usable reads. During NIST's pilot PGM sequencing experiments discussed herein, 31 samples were sequenced concurrently on one 318 chip for each panel. This is roughly half of the maximum number of samples recommended for a 318 chip. In this dataset, two (2) out of 2430 (0.08 %) Identity Panel autosomal SNP genotypes had sequencing coverage below 300 X. The number of Identity Panel Y-chromosome LISNPs with coverage below 150 X was one (1) out of 597 (0.17 %). In the Ancestry Panel data, 90 SNP genotypes out of a total 4455 (2.02 %) were below the suggested minimum 300 X coverage. Developmental validation criteria discussed in the protocol require that the average coverage for a locus should be within two standard deviations of the mean coverage for all loci. Here the mean value of Identity Panel autosomal IISNP loci is 1747.3 X +/- 531.5 X, making the lower threshold for coverage 684.3 X. For Y-chromosome LISNPs, the mean was 849.5 X +/- 236.4 X, making the lower cutoff 376.6 X. For the Ancestry Panel the mean value was 1063.1 X +/- 354.0 X, making the lower threshold for coverage 355.1 X. Higher sequencing coverage was not considered a risk factor for low confidence measurements. Following this principle, one autosomal locus in the Identity Panel (rs2342747 @ 466.1 X) and one Y-chromosome locus (M479 @ 280.2 X) fell below two standard deviations of the mean, while five Ancestry Panel loci were below the criterion: rs1296819 (284.9 X), rs2306040 (323.9 X), rs12439433 (326.0 X), rs16891982 (332.4 X), rs13400937 (346.4 X). Loci with characteristically low coverage are highlighted with orange circles in **Figure 2** and with orange filled cells in **Table 1**.

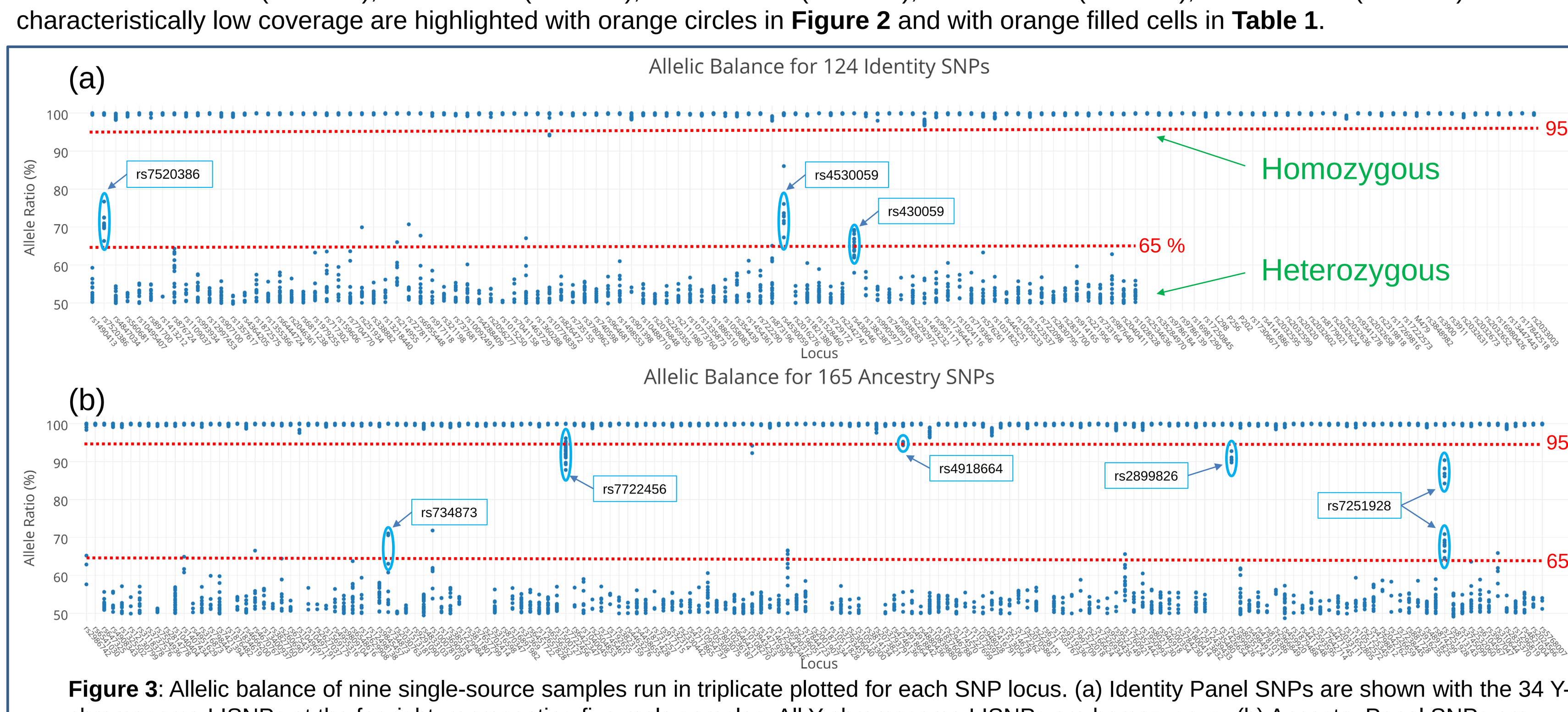


Figure 3: Allelic balance of nine single-source samples run in triplicate plotted for each SNP locus. (a) Identity Panel SNPs are shown with the 34 Y-chromosome LISNPs at the far right, representing five male samples. All Y-chromosome LISNPs are homozygous. (b) Ancestry Panel SNPs are shown. Samples with average allelic balance > 65 % or < 95 % are circled in blue.

Allelic imbalance: heterozygotes should have a sequencing read ratio of 50 % and homozygotes should be 100 %. Deviation from these values may be an indication that a technical issue has occurred with the determination of the genotype. Heterozygotes deviating from the expected 50 % (+/- 15 %) allele ratio may have a SNP underlying a PCR primer binding site which disturbs PCR efficiency or other technical issue such as strand bias or systematic sequencing error. Homozygotes exhibiting a ratio below 100 % (+/- 5 %) could be an indication of systematic sequencing error or bioinformatic sequence alignment issue. Loci with imbalanced allelic ratios are summarized in **Table 3**. A previous study be by Seo *et al*. [7] noted that marker rs4530059 in an earlier version of the Identity Panel displayed allelic imbalance in several heterozygous individuals where the average coverage ratio of A/G was 67.6 %.

Strand bias: defined as the number of plus strand reads divided by the number of minus strand reads, this condition indicates a systematic sequencing error type wherein one strand yields low quality results and is bioinformatically filtered from the final data. Using an arbitrary cutoff of +/- 15 % from the expected ideal ration of 50 % to demark low confidence genotypes, there were five loci in the Identity Panel and nine loci in the Ancestry Panel which yielded at least one data point outside the arbitrary limits of < 35 % and > 65 %. These SNP markers are summarized in **Table 2**. Data points falling outside these arbitrary thresholds may be considered low confidence and will require confirmation with a secondary method.

Discordant replicates: two loci in the Ancestry Panel had discordant genotypes between triplicate data points (see yellow shaded cells in **Tables 1 & 3**). For marker rs3943253 one replicate had a no-call (NN) while the two remaining replicates were homozygous (AA). This marker has an allelic ratio of 93.2 % for the two successfully genotyped replicates and 90.3 for the no-call, however, no quality flag was issued to explain the no-call. All replicates of rs3943253 had some proportion of reads supporting a "G" genotype, which is the alternate allele; the adjacent bases are 5' T and 3' C. The G impurity appears in both forward and reverse reads. No clear explanation is available for this mis-genotype. It may be a case of context-specific sequencing error. A second locus, rs7251928, had one replicate with genotype "CC" and two with "AC". Preliminary data from an alternate platform suggests that the correct call for this locus is "CC". Given that all heterozygotes displayed notable strand bias and allelic imbalance at this locus, resulting genotypes should be considered "very low confidence" until confirmed with an alternate method.

Mendelian inheritance patterns: family trio samples may be used to verify proper genotyping performance. Maternal and Paternal allele inheritance patterns are depicted in **Table 4 below**. Nonconformities in expected inheritance patterns may indicate a technical issue with genotyping a locus. This approach may also be used to identify loci which contain putative SNPs underlying primer binding regions. No inconsistencies in inheritance patterns were identified among family trio samples in the candidate reference materials when evaluated with the Identity Panel or Ancestry Panel. Due to the criteria used in selecting the loci for the Identity Panel (high heterozygosity, low F_{st}) and the Ancestry Panel (low heterozygosity, high F_{st}), there are differences in the percentages of autosomal alleles which may be definitively determined to have originated from one parent versus the other. In the Identity Panel, 38.9 % of loci were informative versus 20.6 % for the Ancestry Panel.

Table 1: Genotypes of nine candidate reference materials characterized for the HID-Ion AmpliSeq Identity Panel (a) and the HID-Ion AmpliSeq Ancestry Panel (b) using the Ion Torrent PGM sequencer.

(a) PGM Identity Panel

Locus	RM 1	RM 2	RM 3	RM 4	RM 5	RM 6	RM 7	RM 8	RM 9
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
rs17222573	-	A	A	-	A	-	A	A	-
M479	-	C	C	-	C	-	C	C	-
rs3848982	-	C	C	-	C	-	C	C	-
rs3900	-	C	G	-	C	-	C	C	-
rs3911	-	A	A	-	A	-	A	A	-
rs2032631	-	G	G	-	G	-	G	G	-
rs2032673	-	T	T	-	T	-	T	T	-
rs2032652	-	T	T	-	T	-	T	T	-
rs16980426	-	T	T	-	T	-	T	T	-
rs13447443	-	A	A	-	A	-	A	A	-
rs17842518	-	G	G	-	G	-	G	G	-
rs2033003	-	A	C	-	A	-	A	A	-

(b) PGM Ancestry Panel

Locus	RM 1	RM 2	RM 3	RM 4	RM 5	RM 6	RM 7	RM 8	RM 9
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
rs7251928	AA	AC	AC	AA	AA	CC	AC	AC	AA
rs8113143	AC	AC	AC	AC	AC	AC	CC	CC	CC
rs3745099	AA	AG	AG	AA	AA	AG	AA	AG	AG
rs2532060	CT	CC	CC	CC	CC	TT	CT	CT	CT
rs6104567	TT	GG	TT	GT	GT	GT	TT	GT	TT
rs3907047	TT	CT	TT	TT	TT	TT	CT	TT	CT
rs310644	TT	TT	CC	TT	TT	TT	CT	CT	TT
rs2835370	TT	TT	TT	TT	TT	TT	CT	CT	TT
rs1296819	CC	CC	AC	CC	AC	CC	AA	AC	AC
rs4821004	CT	CT	CT	CC	CT	CC	CC	CC	CC
rs2024566	AG	AA	AA	AA	AA	AG	AA	AA	AA
rs5768007	CC	CT	CC	CC	CC	CC	CC	CC	CC

Legend	
(yellow)	Discordant Replicate
(orange)	Coverage < 2 SD of the mean
(blue)	Allelic Imbalance

Table 2: Average strand bias (% positive strand) for loci with at least one data point < 35 % or > 65 %.

Identity Panel				Ancestry Panel			
M479 *	17.1	+/-	2.4	rs1040045	39.3	+/-	6.3
rs430046 **	70.1	+/-	11.9	rs1760921	73.6	+/-	8.5
rs1463729	63.8	+/-	2.5	rs1871428	35.5	+/-	2.6
rs2111980	39.3	+/-	3.5	rs1871534	36.5	+/-	2.8
rs4364205	63.0	+/-	2.5	rs2986742	46.7	+/-	8.9
				rs6990312	38.3	+/-	10.2
				rs7251928 **	40.1	+/-	13.3
* Low coverage				rs9845457	45.2	+/-	5.9
** Allelic imbalance				rs12629908	41.6	+/-	4.6

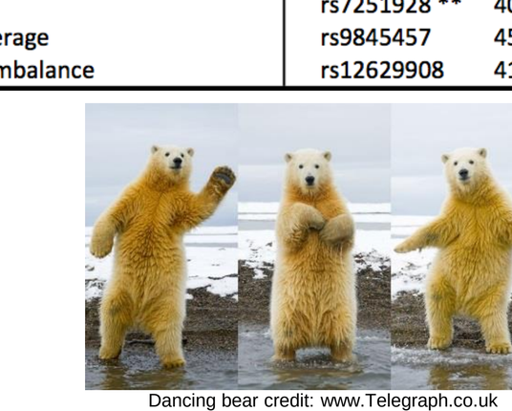

Dancing bear credit: www.Telegraph.co.uk

Table 3: Genotypes and allelic ratios of loci exhibiting allelic imbalance (shaded blue) for the HID-Ion AmpliSeq Identity Panel (a) and HID-Ion AmpliSeq Ancestry Panel (b). * Strand bias observed for this locus.

(a)

Locus	RM 1	RM 2	RM 3	RM 4	RM 5	RM 6	RM 7	RM 8	RM 9
rs7520386	AG	AG	GG	AG	GG	AA	AA	AA	AA
Ratio (%)	71.6 +/- 1.5	70.3 +/- 0.7	99.9 +/- 0.1	71.2 +/- 5.2	99.8 +/- 0.3	99.8 +/- 0.1	99.9 +/- 0.0	99.8 +/- 0.0	99.9 +/- 0.0
rs4530059	GG	AG	AA	GG	GG	GG	AG	AG	GG
Ratio (%)	99.9 +/- 0.1	70.4 +/- 2.9	99.5 +/- 0.1	99.8 +/- 0.2	99.9 +/- 0.1	100.0 +/- 0.0	73.4 +/- 0.5	77.9 +/- 7.4	99.9 +/- 0.1
rs430046 *	CT	CT	CC	TT	CT	CT	CT	CT	CT
Ratio (%)	66.7 +/- 3.6	64.2 +/- 0.4	100.0 +/- 0.0	99.9 +/- 0.0	64.4 +/- 2.3	63.7 +/- 1.6	65.9 +/- 1.7	63.9 +/- 5.5	65.8 +/- 3.7

(b)

Locus	RM 1	RM 2	RM 3	RM 4	RM 5	RM 6	RM 7	RM 8	RM 9
rs734873	GG	GG	GG	GG	GG	GG	AG	AG	AG
Ratio (%)	100.0 +/- 0.0	100.0 +/- 0.1	100.0 +/- 0.0	100.0 +/- 0.1	100.0 +/- 0.0	99.9 +/- 0.0	55.4 +/- 4.7	68.2 +/- 4.5	53.2 +/- 2.7
rs7722456	TT	TT	CC	TT	TT	TT	TT	TT	TT
Ratio (%)	93.5 +/- 1.6	95.2 +/- 1.0	100.0 +/- 0.1	92.8 +/- 0.9	91.6 +/- 0.5	93.3 +/- 1.7	91.7 +/- 2.2	91.3 +/- 3.0	91.3 +/- 1.4
rs3943253	AA	AG	AA	AA	AG	AG	AA	AA	AG
Ratio (%)	99.9 +/- 0.1	54.3 +/- 1.6	93.2 +/- 1.4	99.9 +/- 0.1	51.0 +/- 1.3	51.2 +/- 1.0	99.8 +/- 0.1	99.9 +/- 0.1	51.6 +/- 0.6
rs4918664	AA	GG	GG	AA	AA	AG	AA	AA	AA
Ratio (%)	99.7 +/- 0.4	94.8 +/- 0.5	99.8 +/- 0.2	99.8 +/- 0.2	100.0 +/- 0.1	52.8 +/- 2.6	99.9 +/- 0.1	99.9 +/- 0.0	100.0 +/- 0.0
rs2899826	AA	GG	GG	AA	AA	AA	AA	AA	AA
Ratio (%)	99.8 +/- 0.1	91.3 +/- 0.7	90.4 +/- 0.7	99.8 +/- 0.2	99.8 +/- 0.1	99.9 +/- 0.1	99.9 +/- 0.1	99.9 +/- 0.1	99.9 +/- 0.1
rs7251928 *	AA	AC	AC	AA	AA	CC	AC	AC	AA
Ratio (%)	99.7 +/- 0.2	86.2 +/- 2.0	67.7 +/- 3.3	99.9 +/- 0.1	99.8 +/- 0.1	88.4 +/- 1.8	67.5 +/- 1.0	67.5 +/- 2.5	99.8 +/- 0.1

Discussion:
The majority of SNP loci in the Ion Torrent Identity and Ancestry Panels satisfy proposed requirements for an "informational" classification in a NIST SRM. A small number of loci in these Panels were found to have characteristics signifying potential technical sources of error leading to uncertainty in the measurement. These markers require additional characterization before achieving any status. Confirmation of SNP genotypes with a second NGS method would satisfy the proposed requirements for a "reference value" in a NIST SRM. Final confirmation with Sanger-type sequencing is required to achieve a "certified" value under the proposed framework.

Future Directions:
There are 139 loci which are present in both the HID-Ion AmpliSeq Panels and the Illumina ForenSeq Signature Prep kit. Concordance between the two platforms would qualify those loci as "reference values" for a NIST SRM. Discordant loci will require disambiguation using an alternate method such as Sanger sequencing in order to achieve any status for a NIST SRM. NIST intends to characterize the SNP loci in the Illumina ForenSeq Signature kit. NIST will consider the needs of the forensic community in terms of required levels of confidence in genotypes when prioritizing tasks leading to the potential development of a reference material for SNP genotyping.

Table 4: Mendelian inheritance patterns in Family trio samples for the HID-Ion AmpliSeq Identity Panel SNPs (a) and HID-Ion AmpliSeq Ancestry Panel SNPs (b).

(a)

SNP locus	[illegible]
Father	[illegible]
Mother	[illegible]
Son	[illegible]

(b)

SNP locus	[illegible]
Father	[illegible]
Mother	[illegible]
Son	[illegible]

Legend	
(blue)	Paternal copy
(magenta)	Maternal copy
(grey)	Ambiguous
(white)	Not inherited

Poster available for download from STRBase
http://www.cstl.nist.gov/biotech/strbase/pub_pres/KieslerISFG2015poster.pdf